Estimation de la fonction graphon d'un W-graphe. Application au réseau de la blogosphere politique française

Pierre Latouche, Stéphane Robin

Estimation de la loi a posteriori de la fonction graphon d'un W-graphe. Application au réseau de la blogosphere politique française

Pierre Latouche[1] & Stéphane Robin[2]

[1] *Laboratoire SAMM, EA4543, Université Paris 1 Panthéon-Sorbonne*
[2] *Mathématiques et Informatique appliquées, UMR518, AgroParisTech/INRA, Paris*

Résumé. Les réseaux sont aujourd'hui largement utilisés dans de nombreux domaines scientifiques, en particulier en sciences sociales, afin de représenter les interactions entre entités d'intérêt. Depuis les premiers travaux de Moreno en 1934, de nombreux modèles de graphe aléatoire ont été proposés dans le but d'extraire des informations pertinences à partir de ces données structurées. Le modèle à blocs stochastiques, stochastic block model (SBM) en anglais, permet par exemple de rechercher des groupes de noeuds ayant des profiles de connexions homogènes. Nous nous intéressons ici au modèle de W-graphe qui présente l'intérêt de généraliser la plupart des modèles de graphe aléatoire existants mais pour lequel peu de méthodes existes pour réaliser l'inférence du modèle sur données réelles. Dans un premier temps, nous rappelons comment le modèle SBM peut être représenté sous la forme d'un W-graphe avec une fonction graphon bloc-constante. A l'aide d'un algorithme de type variational Bayes expectation maximization, nous approchons ensuite la loi *a posteriori* des paramètres d'un modèle SBM et nous montrons comment l'incertitude sur les paramètres, caractérisée par cette approximation variationnelle, peut être intégrée de manière analytique afin d'obtenir une estimation de la loi *a posteriori* de la fonction graphon du W-graphe. Dans ce cadre Bayésien, nous dérivons également l'expression de la probabilité d'occurrence d'un motif permettant de tester si un motif est présent de manière exceptionnelle dans un réseau. Ces travaux sont testés sur données simulées et sur un extrait du réseau de la blogosphere politique française.

Mots-clés. Réseau, approximation variationnelle, W-graphe, graphon

Abstract. Networks have been widely used in many scientific fiels, and in particular in social sciences, in order to represent interactions between objects of interest. Since the earlier work of Moreno in 1934, many random graph models have been proposed to extract knowledge from these structured data sets. For instance, the stochastic block model (SBM) allows the search of groups of vertices sharing homogeneous connection profiles. In this work, we consider the W-graph model which is known to generalize many random graph models but for which very few methods exist to perform inference on real data. First, we recall that the SBM model can be represented as a W-graph with a block-constant graphon function. Using a variational Bayes expectation maximization algorithm, we then approximate the posterior distribution over the model parameters

of a SBM model and we show how this variational approximation can be integrated in order to estimate the posterior distribution of W-graph graph function. In this Bayesian framework, we also derive the occurrence probability of a motif. In practice, this allows to test if a motif is over-represented in a given network. All the results presented here are tested on simulated data and the French political blogosphere network.

Keywords. Network, variational approximation, W-graph, graphon function

1 Inférence de la fonction graphon d'un W-graphe

Modèle à blocs stochastiques. Le modèle à blocs stochastiques, stochastic block model (SBM) en anglais, fait l'hypothèse que les n noeuds d'un réseau sont répartis dans Q classes latentes suivant les proportions $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_Q)$. L'association des noeuds aux classes est décrite à l'aide de vecteurs binaires tirés à partir d'une loi multinomiale $\mathcal{M}(1; \boldsymbol{\alpha})$. Les connexions entre noeuds sont alors générées $X_{ij}|Z_i, Z_j \sim \mathcal{B}(\pi_{Z_i,Z_j})$ à partir de lois de Bernoulli dont les paramètres sont caractérisés par la matrice $Q \times Q$ de connectivité $\boldsymbol{\pi} = [\pi_{q\ell}]$, où $\pi_{q\ell}$ est la probabilité de connexion entre un noeud de la classe q et un noeud de la classe l. Par la suite, nous notons $\mathbf{Z} = \{Z_i\}$ l'ensemble des vecteurs latents, $\mathbf{X} = \{X_{ij}\}$ la matrice d'adjacence décrivant l'ensemble des arêtes, et $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\pi})$ l'ensemble des paramètres du modèle.

Lien entre le modèle SBM et les W-graphes. SBM correspond au cas où la fonction graphon W est bloc-constante, avec des blocs rectangulaires de taille $\alpha_k \times \alpha_\ell$ et de hauteur $\pi_{q\ell}$. Plus précisément, si nous notons $\sigma_q = \sum_{j=1}^{q} \alpha_j$ les proportions cumulées, et définissons la fonction

$$C_{\boldsymbol{\alpha}}(u) = 1 + \sum_{q=1}^{Q} \mathbb{I}\{\sigma_q \leq u\},$$

en fixant

$$W(u, v) = \pi_{C(u),C(v)}, \tag{1}$$

le modèle de W-graphe associé correspond alors à un modèle SBM de paramètres $(\boldsymbol{\alpha}, \boldsymbol{\pi})$.

1.1 Algorithme Bayes variationnel pour l'inférence dans SBM

L'inférence du modèle SBM a reçu beaucoup d'attention ces dix dernières années. La principale difficulté vient du fait que la distribution conditionnelle des labels $\mathbf{Z}$ sachant les observations $\mathbf{X}$ n'est pas factorisable et ne peut pas s'écrire sous forme analytique. Des approches de type Monte-Carlo et d'autres basées sur des approximations variationnelles ont donc été proposées. Dans cette présentation, nous considérons l'algorithme variational Bayes expectation maximization (VBEM) de [3] qui permet de traiter efficacement des grands graphes et qui construit une approximation de la loi *a posteriori* des paramètres et

des variables cachées $\mathbf{Z}$ (notées $\widetilde{p}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ et $\widetilde{p}_{\mathbf{Z}}(\mathbf{Z})$). Nous rappelons que cette approximation est obtenue à travers la maximisation par rapport à $\widetilde{p}_{\boldsymbol{\theta}}$ et $\widetilde{p}_{\mathbf{Z}}$ de la fonctionnelle

$$\mathcal{L} = \log P(\mathbf{X}) - KL(\widetilde{p}_{\boldsymbol{\theta}}(\boldsymbol{\theta})\widetilde{p}_{\mathbf{Z}}(\mathbf{Z})||P(\boldsymbol{\theta}, \mathbf{Z}|\mathbf{X})), \tag{2}$$

où KL correspond à la divergence de Küllback-Leibler. En considérant des lois *a priori* conjuguées (i.e. Dirichlet pour $\boldsymbol{\alpha}$ et Beta pour $\pi_{q\ell}$) pour les paramètres du modèle, l'algorithme VBEM approche la loi *a posteriori* à l'aide des distributions suivantes:

$$\begin{aligned} \boldsymbol{\alpha}|\mathbf{X} &\sim \text{Dir}(\mathbf{a}) \qquad \text{where } \mathbf{a} = (a_1, \ldots, a_Q), \\ \pi_{q,\ell}|\mathbf{X} &\sim \text{Beta}(\eta_{q,\ell}, \zeta_{q,\ell}). \end{aligned} \tag{3}$$

Les expressions de a_q, $\eta_{q\ell}$ and $\zeta_{q\ell}$ sont données dans [3].

Nous définissons maintenant une approximation de la loi *a posteriori* de la fonction graphon d'un W-graphe aux coordonnées (u, v). Dans ce but, nous intégrons (1) par rapport aux distributions *a posteriori* approchées de $\boldsymbol{\pi}$ et $\boldsymbol{\alpha}$.

Proposition 1 *Pour un couple $(u, v) \in [0, 1]^2$, $u \leq v$, en utilisant un modèle SBM à Q classes, l'approximation Bayésienne variationnelle de $W(u, v)$ est $\widetilde{p}(w|\mathbf{X}, Q) =$*

$$\sum_{q \leq \ell} b(w; \eta_{q,\ell}, \zeta_{q,\ell}) \left[F_{q-1,\ell-1}(u, v; \mathbf{a}) - F_{q,\ell-1}(u, v; \mathbf{a}) - F_{q-1,\ell}(u, v; \mathbf{a}) + F_{q,\ell}(u, v; \mathbf{a})\right]$$

où

- $\mathbf{a}, \eta$ *and* ζ *sont les hyperparamètres obtenus par l'algorithme VBEM;*
- $b(\cdot; \eta, \zeta)$ *est la densité de probabilité de la loi Beta*(η, ζ)*;*
- $F_{q,\ell}(u, v; \mathbf{a})$ *est la fonction de répartition jointe de (σ_q, σ_ℓ) où $\boldsymbol{\alpha}$ suit une loi de Dirichlet Dir*$(\mathbf{a})$*.*

Dans cette présentation, nous décrirons la preuve de ce résultat avant d'utiliser la méthode sur des données simulées et réelles. Un point essentiel de l'approche est que la fonction de répartition jointe $F_{q,\ell}$ peut être calculée par un algorithme récursif décrit dans [5].

L'estimateur approché de l'espérance *a posteriori* se déduit directement de la proposition précédente: $\widetilde{\mathrm{E}}[W(u, v)|\mathbf{X}] =$

$$\sum_{q \leq \ell} \frac{\eta_{q,\ell}}{\eta_{q,\ell} + \zeta_{q,\ell}} \left[F_{q-1,\ell-1}(u, v; \mathbf{a}) - F_{q,\ell-1}(u, v; \mathbf{a}) - F_{q-1,\ell}(u, v; \mathbf{a}) + F_{q,\ell}(u, v; \mathbf{a})\right].$$

L'écart type cette estimation du graphon se calcule également de manière analytique.

2 Quelques résultats expérimentaux

Notre présentation sera composée de deux parties, la première traitant, comme décrit précédemment, de l'approximation de la fonction graphon d'un W-graphe, la deuxième se concentrant sur le calcul de la probabilité d'occurrence d'un motif dans le cadre Bayésien considéré. Nous avons seulement présenté dans ce papier les résultats de la première partie et nous donnons donc dans cette section les résultats expérimentaux associés. Ces résultats ont été obtenus en analysant un extrait du réseau de la blogosphere politique française où chaque noeud représente un blog et les arêtes caractérisent les hyperliens connus entre les blogs.

Bibliographie

[1] P. Bickel, A. Chen (2009), A non parametric view of network models and newman-girvan and other modularities. In Proceedings of the National Academy of Sciences, 106, 21068-21073.
[2] J.J. Daudin, F. Picard, and S. Robin (2008), A mixture model for random graphs, Statistics and Computing, 18(2), 173-183.
[3] P. Latouche, E. Birmelé, and C. Ambroise (2012), Variational Bayesian inference and complexity control for stochastic block models, Statistical Modelling, 12(1), 93-115.
[4] F. Picard, J.-J. Daudin, M. Koskas, S. Schbath, and S. Robin (2008), Assessing the exceptionality of network motifs, Journal of Computational Biology, 15(1), 1-20.
[5] H. Exton (1976), Multiple hypergeometric functions and applications, Mathematics and its applications, Ed John Wiley.

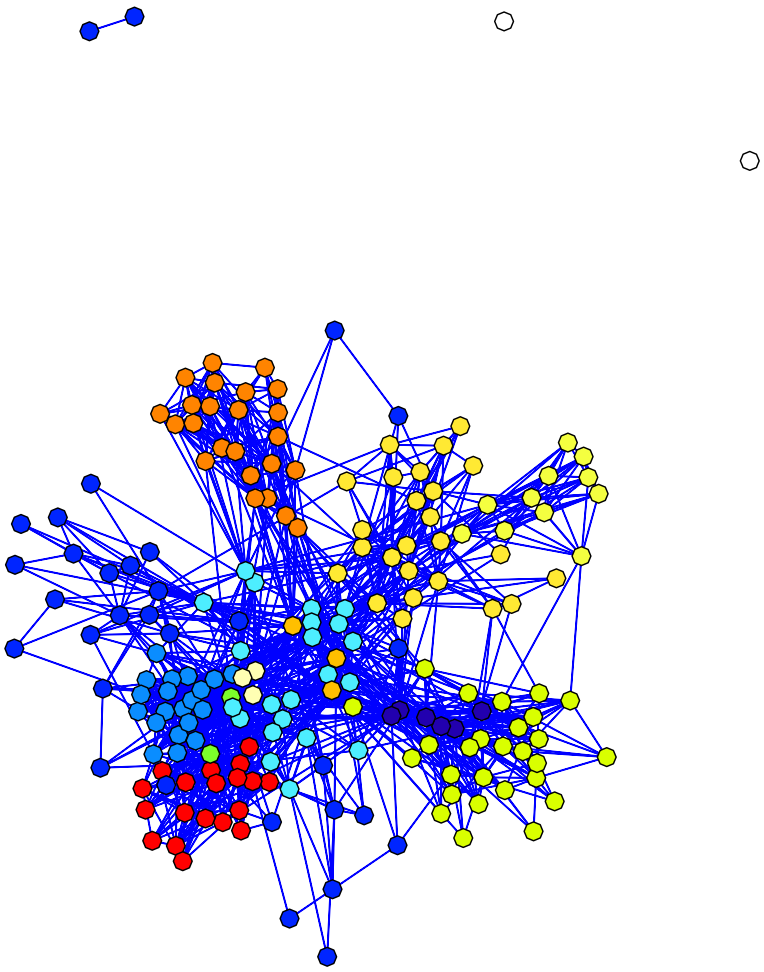

Figure 1: Classification des noeuds du réseau de la blogosphere politique française à l'aide de l'algorithme VBEM pour SBM.

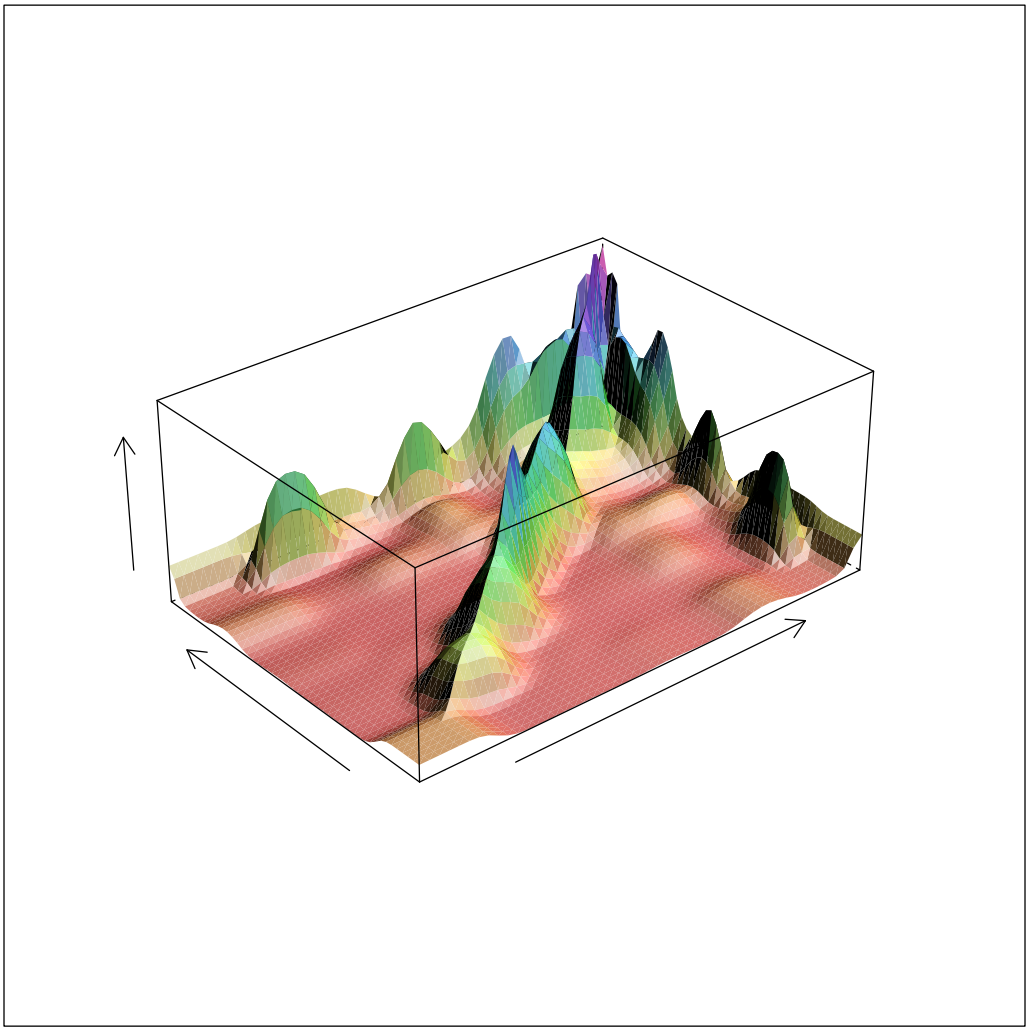

Figure 2: Estimation de la fonction graphon du réseau de la blogosphere politique française.